Empirical Flow Parameters: A Tool for Hydraulic Model Screening

Theodore G. Cleveland[1], Caroline M. Neale[2], Cristal C. Tay[2], George R. Herrmann[3], and

[1]Associate Professor, [2]Graduate Research Assistant, [3]Post-Doctoral Researcher
Department of Civil, Environmental, and Construction Engineering, Texas Tech University, 10th and Akron, Lubbock, TX 79409-1023; PH (806) 834-5101; email: theodore.cleveland@ttu.edu, cristal.tay@ttu.edu, caroline.neale@ttu.edu, ghermann@suddenlink.net.

ABSTRACT

Conditional distributions constructed from an extensive database are presented as an alternative to regression equations to estimate mean section velocity or other selected hydraulic parameters for storm flows or other conditional discharges at un-gauged locations.

Illustrative examples are presented showing how to generate the distributions in the **R** programming environment and represent them as graphs, tables, or quantile functions.
A particularly powerful feature of **R** as the tool to access the database is the ability to rapidly construct conditional distributions, where the distributional information is conditioned on some other criteria in the database. Conditioning addresses considerations such as the 95th percentile discharge from all observations being far less meaningful than the 95th percentile discharge for observations from drainage areas less than 40 square miles (Discharge conditioned on drainage area). Several other conditioning examples are presented.

The use of the tool as a screening instrument for hydraulic modeling is discussed.

INTRODUCTION

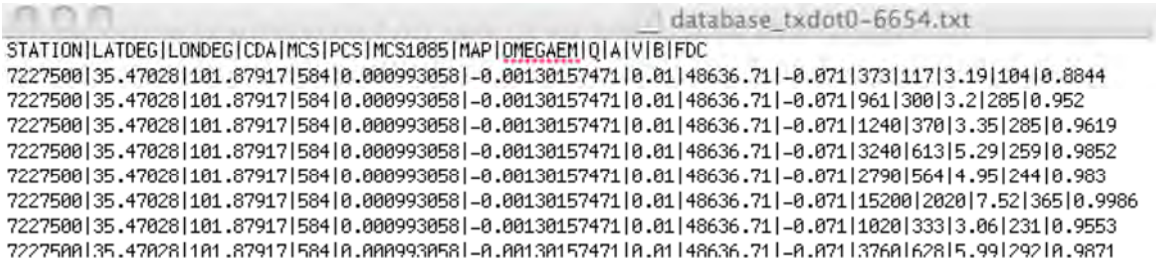
Empirical flow parameter distributions described herein are a statistical tool to estimate mean section velocity or other selected parameters for storm flow or other conditional discharges at ungauged locations in Texas. These distributions are an alternative to a regional regression or regression-like approach (Asquith, Herrmann, and Cleveland, 2013) that provides an equation for estimation of the expected value for mean velocity and/or discharge for an ungauged location and the prediction limits of that estimate.

The empirical distributions presented herein are based on data retrieved from the NWIS (USGS, 2009). Accessing the distributions is accomplished using **R** (R Development Core Team (2011)), and examples of such use derived from Cleveland and others (2013) are provided herein.

A particularly powerful feature of **R** as the tool to access the database is the ability to rapidly construct conditional distributions, where the distributional information is conditioned on some other criteria in the database. Conditioning addresses concerns such as where the 95th percentile discharge from all observations may be less meaningful than the 95th percentile discharge for all observations from drainage areas less than 40 square miles (Discharge conditioned on drainage area).

METHODOLOGY

The database is an ASCII text file that about 87,000 records from various gaging stations in Texas. Figure 1 is a screen capture of the first few rows of the database. The data are arranged in columns using the pipe symbol "|" as the delimiter.


database_txdot0-6654.txt

```
STATION|LATDEG|LONDEG|CDA|MCS|PCS|MCS1085|MAP|OMEGAEM|Q|A|V|B|FDC
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|373|117|3.19|104|0.8844
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|961|300|3.2|285|0.952
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|1240|370|3.35|285|0.9619
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|3240|613|5.29|259|0.9852
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|2790|564|4.95|244|0.983
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|15200|2020|7.52|365|0.9986
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|1020|333|3.06|231|0.9553
7227500|35.47028|101.87917|584|0.000993058|-0.00130157471|0.01|48636.71|-0.071|3760|628|5.99|292|0.9871
```

Figure 1. Screen Capture first few rows of database

The column headings in the figure correspond to the following descriptions:

1. `STATION` is 8-digit USGS station identification code. These codes can be entered into the NWIS web interface to recover textural description of the particular station and other locational information.
2. `LATDEG` is the latitude in degrees and decimal degrees (dd.ddddd).
3. `LONDEG` is the longitude in degrees and decimal degrees (ddd.ddddd).
4. `CDA` is the contributing drainage area to the station in square miles (mi2).
5. `MCS` is the main channel slope (Asquith and Slade, 1997).
6. `PCS` is the proximal channel slope.
7. `MCS1085` is the 10-85 main channel slope (Gordon and others, 2004).
8. `OMEGAEM` is the OmegaEM parameter (Asquith, and Roussel , 2009).
9. `Q` is the observed discharge in cubic feet per second (ft^3/s).
10. `A` is the cross sectional flow area (at the above observed discharge) in square feet (ft^2).
11. `V` is the mean section velocity (ratio of discharge to flow area) in feet per second (ft/s).
12. `B` is the topwidth (at the observed discharge) in feet (ft).
13. `FDC` is the flow duration curve for the associated station. There are hundreds of stations represented in the database, and a flow duration curve was computed for each station. The exceedance probabilities of flow for that station were maintained to provide a useful conditioning capability.

Unconditioned distributions are empirical distributions based on the entire database without regard to any of the other retained variables. They are constructed using either the `quantile()` function in **R** or using a Weibull plotting position formula. Loading the database into **R** is illustrated in Listing 1. Several approaches are illustrated in using the database to generate useful plots, tabulations, and finally direct access using the `quantile()` function.

Listing 1. R code loading the database and preparing some useful plot labels.

```
# EXAMPLE 1 # ** Loading the database into R, a useful function, and persistent plot labels
txdot0_6654.db <- read.table("database_txdot0-6654.txt",
                             header=TRUE, sep="|")
DB <- txdot0_6654.db; # shorten the database name considerably

"weibullpp" <- function(x, sort=TRUE) {
    denom <- length(x) + 1
    ranks <- rank(x, ties.method = "first")
    ifelse(sort, return((sort(ranks))/denom), return((ranks)/denom))
}
XLAB  <- "NONEXCEEDANCE PROBABILITY"
YLABV <- "VELOCITY, IN FEET PER SECOND"
YLABQ <- "DISCHARGE, IN CUBIC FEET PER SECOND"
```

The entire database can now be queried to produce an empirical discharge distribution for discharge for the measured values in Texas. Figure 2 results from computing the plotting position for each entry in the database.

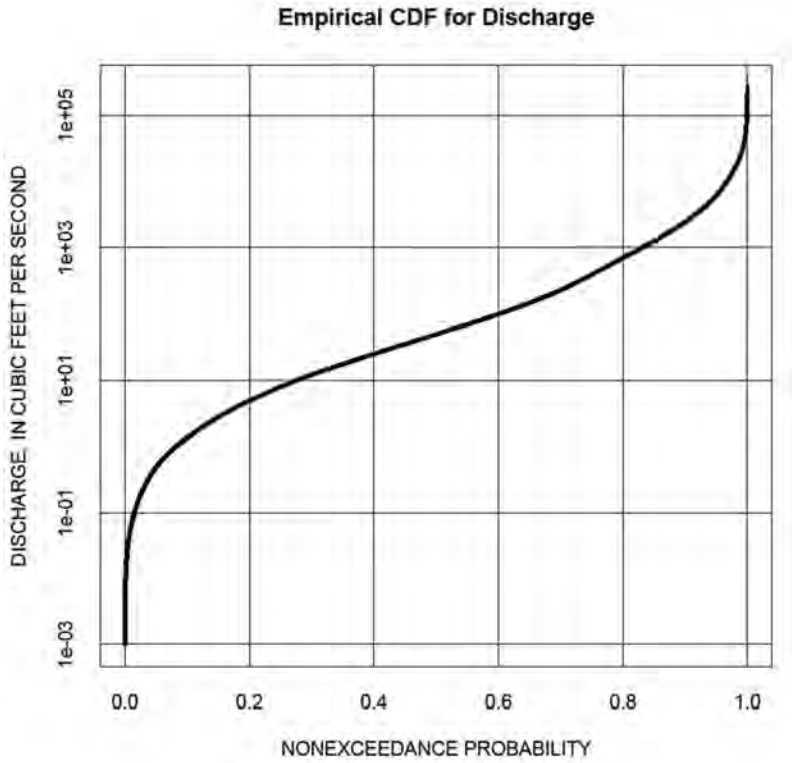


Figure 2. Empirical Cumulative Discharge Distribution

The script in **R** for generating the figure is shown in Listing 2.

Listing 2. **R** script to generate CDF of Discharge. Other variables in the database are addressed in a similar fashion.

```
# EXAMPLE 2 #  ** Empirical Unconditioned Distribution of Discharge
plot(weibullpp(DB$Q),sort(DB$Q),log="y",xlab=XLAB,ylab=YLABQ,type="s",lwd=3,tck=1, main="
     Empirical CDF for Discharge")
```

Table 1 is a tabular representation of the curve displayed in Figure 2.

Table 1. Tabular CDF of discharge

Percent Non-Exceedance	Discharge (cfs)
0.0000*	0.001
0.0001	0.003
0.0010	0.010
0.0100	0.060
0.0500	0.480
0.1000	1.360
0.2000	4.950
0.3000	12.200
0.4000	25.100
0.5000	49.800
0.6000	99.300
0.7000	227.000
0.8000	701.000
0.9000	2440.000
0.9500	6160.000
0.9900	26800.000
0.9990	79168.200
0.9999	147988.700
1.0000**	268600.000

* Smallest observed value in the database.
** Largest observed value in the database.

The script in **R** for generating the table in Table 1 is shown in Listing 3. The tabular output in **R** was copied into an Excel worksheet, then pasted into a typesetting program to generate Table 1. This step (cut-paste-reformat) is not needed to use the tools, but was used for this particular table to highlight that the `quantile()` function returns the smallest value in the database at the 0^{th} percentile and the largest value at the 100th percentile level.

Both these representations (the empirical distribution plot and the table) can be used to answer questions like: What is the probability of observing a discharge less than 12.2 cubic feet per second?[1] The answer would be to find the value of interest and then read the associated non-exceedance probability either from the graph or the tabulation. In this case about 30 percent of the observations are smaller than 12.2, so one could expect to observe such a discharge in a random measurement about 30 percent of the time.

[1] Without regards to where in the state we may be.

Lastly, instead of using a chart or a tabulation, the result can be recovered directly from the database using the `quantile()` function in R. Listing 4 illustrates the use of R to directly locate a value based on a desired non-exceedance probability.

Listing 3. R script to generate tabular CDF of discharge. Other variables in the database are addressed in a similar fashion.

```
# EXAMPLE 3 #  ** Empirical Unconditioned Distribution of Discharge -- Tabular
    Representation
# Useful quantiles for tabular presentations
> EMPQUANT<-c
    (0,0.0001,0.001,0.01,0.05,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95,0.99,0.999,0.9999,1.0)

> cbind(quantile(DB$Q,EMPQUANT)) # output from R below
                [,1]
0%      1.000000e-03
0.01%   2.715900e-03
0.1%    1.000000e-02
1%      6.000000e-02
5%      4.800000e-01
10%     1.360000e+00
20%     4.950000e+00
30%     1.220000e+01
40%     2.510000e+01
50%     4.980000e+01
60%     9.930000e+01
70%     2.270000e+02
80%     7.010000e+02
90%     2.440000e+03
95%     6.160000e+03
99%     2.680000e+04
99.9%   7.916820e+04
99.99%  1.479887e+05
100%    2.686000e+05
>
```

Listing 4. R script using `quantile()` function to approximate non-exceedance values of variable in the database.

```
# EXAMPLE 4 #  ** Empirical Unconditioned Distribution of Discharge -- Quantile Function to
    Assess a Particular Value
> quantile(DB$Q,0.3) #  output from R below
 30%
12.2
```

The real value of the database and empirical distributions accessed using **R**, is the ability to condition the distributions on other variables in the database. This conditioning is presented by example, but in essence is a filtering process.

A logical conditioning is to ask from the database what is a certain non-exceedance probability associated with all discharges recorded from drainage areas less than some value (or even bracketed). This question is a conditioned probability statement. Operationally we would search the database and exclude all records associated with drainage areas larger than the prescribed conditioning value, then compute the empirical distribution from the remaining values.

As an example consider what is the empirical non-exceedance discharge distribution for gages having contributing drainage areas of 100 square miles or less? The drainage area is the variable CDA and building the conditional distribution is illustrated in Listing 5.

Listing 5. **R** script to generate conditional distributions of one variable conditioned on value of another variable.

```
# EXAMPLE 9
#   ** Empirical Distribution of Discharge Conditioned on Contributing Drainage Area
# Get all measurements for watershed area less than 100 square miles
> QQ<-DB[DB$CDA < 100,]$Q  # Filter the database, put results in QQ
# Plot an empirical distribution (sneaky here, actually don't use QQ just yet!)
> plot(weibullpp(DB[DB$CDA < 100,]$Q),sort(DB[DB$CDA < 100,]$Q),log="y",xlab=XLAB,ylab=
    YLABQ,type="s",lwd=3,tck=1, main="Conditional Empirical CDF for Discharge for CDA < 100
     sq.mi.")
# Compute the 50th percentile using the quantile() function of R---this is the median
> print(quantile(QQ,0.5)) # output from R below
  50%
7.735
# Generate a tabulation
> cbind(quantile(QQ,EMPQUANT)) # output from R below
               [,1]
0%      1.00000e-03
0.01%   1.37310e-03
0.1%    1.00000e-02
1%      4.00000e-02
5%      1.70000e-01
10%     4.10000e-01
20%     1.12000e+00
30%     2.24000e+00
40%     4.19400e+00
50%     7.73500e+00
60%     1.50000e+01
70%     3.07000e+01
80%     6.82000e+01
90%     2.46000e+02
95%     8.29350e+02
99%     3.70380e+03
99.9%   1.02807e+04
99.99%  1.53000e+04
100%    1.65000e+04
>
```

The result of conditioning is apparent in the median value of discharge. When all drainage areas were considered, the median discharge was about 50 cfs, but when conditioned on contributing drainage area less than 100 square miles, the median is around 8 cfs – about six times smaller. The result is anticipated. Smaller drainage areas should produce smaller discharges for similar weather conditions. As a guideline, the authors suggest that when conditioning, the analyst check the array sizes and try to maintain about 100 records after conditioning; with this suggestion each retained record represents about 1 percent of any empirical distribution subsequently generated.

Multiple conditioning based on several variables is feasible. Listing 6 is an example of a multiple conditioning empirical distribution where the analyst seeks the 95th percentile of discharges from contributing drainage areas less than 100 square miles, with topwidth less than 40 feet, and discharges greater than the 80th percentile on the station's individual flow duration curve.

Listing 6. R script for estimating the non-exceedence value of a variable conditioned on the values of multiple other variables.

```
# EXAMPLE 10 #   ** Empirical distribution of discharge for multiple conditions
# Get all measurements for watershed area less than 100 square miles
#   and topwidth less than 40 feet
#   and greater than 80th percentile on the flow-duration curve
# Compute the 75th percentile of these measurements
> print(quantile(DB[DB$CDA < 100 & DB$B < 40 & DB$FDC > 0.80, ]$Q, 0.95)) # output from R
    below
  95%
224.6
```

The result in this case is a discharge of about 225 cfs.

RESULTS

The utility of multiple conditioning provides a tool to screen hydraulic model results. The use of the database as a screening tool is illustrated by example.

For example suppose the analyst seeks the empirical velocity distribution for drainage area less than 10 square miles, topwidth less than 30 feet, and flows (in the retained records) that are at the 90th percentile or greater on the individual station flow duration curves. Listing 7 illustrates conditioning the velocity on drainage area, topwidth, and station flow duration curve values. The result is tabulated using the `quantile()` function approach presented in earlier examples.

Listing 7. Multiple conditioning for Mean Section Velocity

```
# EXAMPLE 12 #   ** Empirical Distribution of Velocity Conditioned on Contributing Drainage
     Area, Topwidth, and Flow Duration
# Get all measurements for watershed area less than 10 square miles
#   topwidth less than 30 feet
#   and greater than 90th percentile on the station flow-duration curve
VV<-DB[DB$CDA < 10 & DB$B < 30 & DB$FDC > 0.9,]$V  # filter database, put result in VV
> length(VV) # check length -- OK but getting to be on the small side!
[1] 132
# Build a tabular empirical distribution.
> cbind(quantile(VV,EMPQUANT)) # output from R below
             [,1]
0%      0.340000
0.01%   0.341179
0.1%    0.351790
1%      0.433100
5%      0.530000
10%     0.662000
20%     0.830000
30%     1.006000
40%     1.344000
50%     1.485000
60%     1.886000
70%     2.217000
80%     2.772000
90%     3.453000
95%     4.368000
99%     5.531600
99.9%   5.964630
99.99%  5.996463
100%    6.000000
```

Once the output is examined, the median value for these conditions is about 1.5 feet per second, the largest value retained after the conditioning is 6 feet per second. We can now consider what information the distribution is conveying. If a modeler were to calculate a velocity for a contributing drainage area of 10 square miles, for a topwidth at the point of interest of 30 feet and arrive at a value of say 7 feet per second, the modeler should be concerned. The database suggests that such a value has not yet been observed in Texas streamflow, even when considering flows at the 90th percentile on the individual station flow duration curves. Hence the value of 7 feet per second, unless otherwise explained, would be disturbing.[2]

[2] The value from the model may indeed correct, but based on observations it is unusual. The whole point of the tool is to guide when a value is unusual and help identify potential data entry errors that could otherwise go unnoticed.

The database can also be used to compute ancillary (or derived) values, such as Froude number,

$$Fr = \frac{V}{\sqrt{gA/B}} \quad [1]$$

where V is the mean section velocity, A is the cross sectional flow area, B is the topwidth, and g is gravatitional acceleration.

Listing 8 illustrates the construction of a derived distribution for Froude number for contributing drainage areas less than 10 square miles. The result is tabulated using the `quantile()` function. For this example the median Froude number for such conditions is about 0.25, a decidedly subcritical flow. In fact, based on the database, supercritical flow is unusual occurring above the 99.9th percentile. Supercritical flow is indicated for some measurements, but very few are computed in this flow regime in the database.

Listing 8. Building a derived (ancillary) empirical distribution.

```
# EXAMPLE 13 #   ** Empirical Distribution of Froude number conditoned on drainage area
    less than 10 square miles.
# Get all measurements for watershed area less than 10 square miles, compute the Froude
    number, tabulate the result
FR <-DB[DB$CDA < 10,]$V/sqrt(32.2*DB[DB$CDA < 10,]$A/DB[DB$CDA < 10,]$B)  # compute Fr for
    CDA < 10, put into FR
> cbind(quantile(FR,EMPQUANT)) # output from R below
              [,1]
0%      0.007855307
0.01%   0.008281195
0.1%    0.012114189
1%      0.024655607
5%      0.059134675
10%     0.079623847
20%     0.122351979
30%     0.164806117
40%     0.204549940
50%     0.249712049
60%     0.296877395
70%     0.354723991
80%     0.421251480
90%     0.519535661
95%     0.596747258
99%     0.728635974
99.9%   0.950658157
99.99% 1.211274071
100%    1.240231395
```

SUMMARY

Empirical distributions for certain hydrologic and hydraulic properties from gauging stations in Texas were presented. The underlying database was described and the procedure to load the database into the **R** programming environment was presented. The database itself is an ASCII text file and while it is intended for use with **R**, it could conceivably be loaded into Excel. Additional illustrative examples were presented showing how to generate various distributions in **R** and represent them as graphs, tables, or `quantile()` function calls. Both unconditioned and conditioned distributions were presented. The use of conditioning allows the analyst to select affiliated variables from the database as conditions and retain only those records which satisfy the conditions. These records can subsequently be converted into empirical distributions using either the `weibullpp()` function or the built-in `quantile()` function. Interpreting the results

of multiple conditioning as a way to screen hydraulic model results was presented by example.

REFERENCES

Asquith, W.H., and Slade, R.M., 1997, Regional equations for estimation of peak-stream flow frequency for natural basins in Texas: U.S. Geological Survey Water Resources Investigations Report 96{4307, http://pubs.usgs.gov/wri/wri964307/.

Asquith, W.H., and Roussel, M. S., 2009, Regression equations for estimation of annual peak-stream flow frequency for undeveloped watershed in Texas using an L-moment-based, PRESS-minimized, residual adjusted approach. U.S. Geological Survey Scientific Investigations Report 2009-5087.

Asquith, W.H., Herrmann, G.R., and Cleveland, T.G., 2013, "Generalized Additive Regression Models of Discharge and Mean Velocity generally associated with Direct-Runoff Conditions in Texas: The Utility of the U.S. Geological Survey Discharge Measurement Database" American Society of Civil Engineers, Journal of Hydrologic Engineering, Vol. 18, No. 10, pp 1331-1348

Cleveland, T. G., Strom, K. B., Sharif, H. Liu, X. 2013. "Empirical Flow Parameters – A Tool for Hydraulic Model Validity Assessments" Report 0-6654-1, Texas Department of Transportation.http://www.depts.ttu.edu/techmrtweb/Reports/Complete%20Reports/0-6654-1_Final.pdf

Gordon, N.D., T.A. McMahon, B.L. Finlayson, C.J. Gippel, R.J. Nathan, 2004, Stream Hydrology: An Introduction for Ecologists (second edition). John Wiley, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 429 p.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

U.S. Geological Survey, 2009, Streamflow measurements for Texas: USGS National Water Information System, accessed March 1, 2009, http://waterdata.usgs.gov/tx/nwis/measurements/?site_no=STATIONID&agency_cd=USGS.